FEATURE-BASED VIDEO KEY FRAME EXTRACTION FOR LOW QUALITY VIDEO SEQUENCES

Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora

Communication Systems Group, TU Berlin,
EN-1, Einsteinufer 17, 10587 Berlin, Germany
$\{kelm, schmiedeke, sikora\}$ @nue.tu-berlin.de

ABSTRACT

We present an approach to key frame extraction for structuring user generated videos on video sharing websites (e. g. YouTube). Our approach is intended to link existing image search engines to video data. User generated videos are, contrary to professional material, unstructured, do not follow any fixed rule, and their camera work is poor. Furthermore, the coding quality is bad due to low resolution and high compression. In a first step, we segment video sequences into shots by detecting gradual and abrupt cuts. Further, longer shots are segmented into subshots based on location and camera motion features. One representative key frame is extracted per subshot using visual attention features, such as lighting, camera motion, face, and text appearance. These key frames are useful for indexing and for searching similar video sequences using MPEG-7 descriptors [1].

1. INTRODUCTION

Temporal video segmentation is the first step towards automatic annotation of digital video sequences. Its goal is to divide the video stream into a set of meaningful segments that are used as basic elements for indexing and classification. From the indexing point of view, it is easier to index a few frames from different shots of a video instead of indexing all the frames. Therefore, shot detection is a useful first step.

Cotsaces et al. give an overview of the different types of shot changes in [2]. Lienhart [3] tests the performance of various existing shot detection algorithms on a diverse set of video sequences and concludes that all algorithms are been adversely affected by motion in the video. Seaz et al. [4] rely on two different video characteristics, edges and luminance, to detect scene changes. Won et al. [5] and Qian et al. [6] develop mathematical models for dissolves and fades. Their detection method is based on the difference of modelling errors to an ideally modelled transition. Dissolves [5] are detected using the parabola characteristics of the luminance variance curve and fades [6] are detected using the accumulated histogram difference. Although existing research on shot boundary detection is active and extensive, most approaches focus on broadcast material or otherwise professional recordings. In contrast, our framework focuses to detect shot changes in low quality user generated web videos. There are several challenges applying conventional shot detection algorithms. User generated videos lack image sharpness in contrast to professional recordings with the same resolution. The images are mostly blurred due to their high compression rate. Additionally, these video sequences are rarely filmed with a steady hand, so the pictures are shaky. Also, the structures differ completely. There are lots of video which are unedited, so they contain only a single shot or a few shots separated by hard cuts. On the other hand, there are edited video sequences containing frequent special effect transitions in all kinds.

Next useful step for indexing is the selection of meaningful frames to generate video summaries. The key frame extraction of Zhao and Cai [7] is based on visual attention and affective models which fuse film elements such as lighting and camera motion. In [8], the autoregressive prediction error is used to find key frames by selecting frames with the smallest prediction error in the shot. Sun et al. [9] assume key frames at the peaks of the distance curve of colour distribution between frames in the shot and a "'temporally maximum occurrence frame"'. An information theory based approach is used by Cernekov et al. [10] to detect shots and extract key frames using mutual information and the joint entropy.

We face these challenges of detecting shots and extracting key frames in low quality web videos.

2. SEGMENTATION FRAMEWORK

Our framework segments user generated web video in shots using algorithms proposed in section 2.1. Since the quality of these web videos is low, only a few features are left available for the shot change detection. Edge and motion features are not useful of this kind of video sequences, because the image sharpness and the camera work are too bad for a reliable analysis. Consequently, we rely on colour features to de-

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 216444.

tected shot changes. Shaky pictures are also the reason why algorithms based on single pixels and small blocks in spatial domain cannot be applied. On the other hand averaging over too large regions might fail to divide similar shots. We use different spatial domains according to the characteristics of transitions.

Since user generated video sequences are sometimes unedited and have longer shots, we introduce a method to subdivide them into shorter subshots (Fig. 1) in section 2.2.

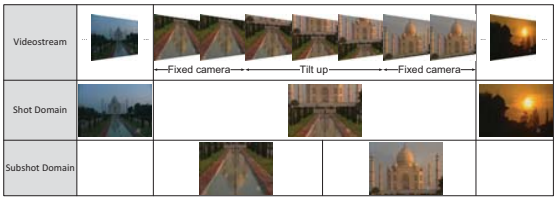

Fig. 1. Segmentation and key frame extraction for the best representation of a shot and subshot

2.1. Shot boundary detection

Our approach detects independently the three most commonly used transitions in web videos; hard cut, fade and dissolve. Also the methods are well-known in this subsection; there are some improvements in detail according to the particular videos processed here.

2.1.1. Hard cut

The hard cut is the most important transition, especially for unedited video material, and describes the abrupt change between two shots. We use colour histograms to detect hard cut in a way similar to [3], except that the spatial correlation is taken into account by dividing each frame in $A = 3$ horizontal areas without getting sensitive to small camera and object motion. For noise robustness, we reduce the number of histogram bins in HSV colour spaces to 256 bins according to MPEG-7 Scalable Colour (SCD) [1]. The discontinuity D_a between frames is evaluated by the L_2-norm within a temporal window over $N = 5$ frames. An adaptive threshold $th_a(n)$ is calculated independently for each area:

$$th_a(n) = \alpha \cdot \left[\left(\sum_{m=n-N}^{n+N-1} D_a(m, m-1) \right) - D_a(n, n-1) \right] + \beta \quad (1)$$

The constants regulate the adaptive thresholds, in empirical tests $\alpha = 2.5$ and $\beta = 8.7$ lead to the best result. A hard cut is detected, if $D_a(n, n-1) > th_a(n)$ and $D_a(n+1, n) < th_a(n+1)$ is true for all areas ($a = 1 \ldots 3$).

2.1.2. Fade

The disappearing shot fades into a blank frame followed by the appearing shot that fades in. The luminance average Y_μ and variance Y_{σ^2} exhibit a certain pattern which is insensitive to noise and camera motion. The approach suggested by [3] also works well for the low-quality videos. So we detect the centre of fades by thresholding the strictly monotonic decreasing first derivative of the luminance variance.

2.1.3. Dissolve

A dissolve is defined by a temporal overlap of a few frames of the disappearing and appearing shot. The variance of luminance during the overlap progresses on a parabola with a minimum at the centre of the transition (Fig 2). Dissolve candidates are extracted by the characteristics of the first and second derivatives of Y_{σ^2} [5]. Due to low image quality and fast camera operations, this approach produces many dissolve candidates. Our efforts lie in the verification of these candidates. The extrema of the second derivative mark the start and end point of a dissolve candidate which is used for generating an ideally modelled dissolve. The verification is done by

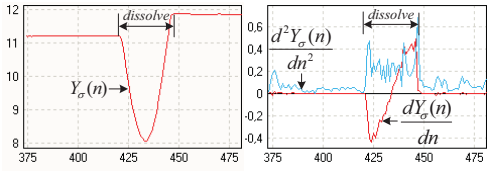

Fig. 2. Standard deviation curve (left), first and second derivatives of the standard deviation curve (right)

considering following aspects. The cross-correlation between the course of the first derivatives of the candidate region and the ideal dissolve must exceed the threshold $th_{cc} = 0.9$. Additionally, the mean of first derivative must exceed a small positive threshold.

2.2. Subshot boundary detection

The segmentation of shots in smaller units has a special importance for single shot videos and for long shots. This division into subshots targets to detect new visual content appearing through camera operations or special effects. The candidates for subdividing are determined by a location memory model and the camera motion.

The location memory model is based on accumulated colour histograms H in the HSV colour space (with K bins). At each point in time, this model computes the normalised distance $dist_{loc}(n)$ to an average histogram of the previous $N = 480$ frames ($T = 16$ s):

$$dist_{loc}(n) = \sum_{k=0}^{K-1} \left| H_n(k) - \frac{1}{N-1} \sum_{i=n-N}^{n-1} H_i(k) \right|. \quad (2)$$

It can be assumed that location or lighting changes come along with high values in the distance function. Consequently, special effect transitions are detected by comparing $dist_{loc}(n)$ to the threshold $th_{loc} = 60\%$.

The camera motion is estimated using the motion vectors from the compressed video stream. Each macroblock has a motion vector which points to a similar block in a reference frame. These vectors are a good indicator for camera motion. We use median filtering in order to take into account shaky camera work and low image quality. The next step is the partitioning of the frame in four areas and computing the respective average motion vector. The angle and the intensity of each vector are compared to a template according to Fig. 3. The motion vectors point to or originate from the centre during the camera operation zoom or dolly. Besides, progressive camera motions are recognized by the detection of four similar angles. For every template, there is a threshold $th_{cm} = 20°$ for the deviation of the angle. Temporal outliers are removed

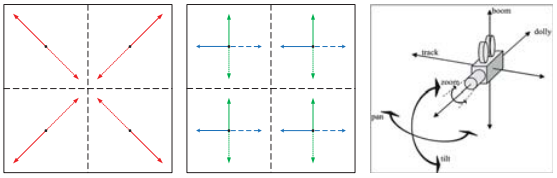

Fig. 3. Camera operations (right) [1] and their detection templates: zoom/dolly (left), track/pan and tilt/boom (middle)

by median filtering. As new visual content can only appear through camera operations or undetected transitions, a fixed camera frame after a panning operation is a candidate for subshot boundary.

A candidate becomes a boundary to divide the shot, if the candidate frames differ visually. The visual similarity between two candidate frames is measured using SCD [1] and a threshold of 33%.

2.3. Extraction of key frames

Our key frame extraction approach targets to get a representative frame of a (sub)shot with noticeable visual content and in best possible quality. The motion intensity f_{mi} calculated over all $M \times N$ motion vectors in compressed domain is proportional to blurring effects in each frame:

$$f_{mi}(n) = \sqrt{\left(\frac{1}{M}\sum_{i=0}^{M-1} mv_x(i)\right)^2 + \left(\frac{1}{N}\sum_{j=0}^{N-1} mv_y(j)\right)^2}. \quad (3)$$

Consequently, less noisy (key) frames of each subshot can be found at the minimum of the intensity curve.

Camera motion is an important tool for the director of a video. "Zooming in" attracts attention to details while "zooming out" emphasizes the surroundings. In addition to this, a fast panning or tilting camera indicates unimportant scene fragments. The camera motion f_{cm} is defined as:

$$f_{cm}(n) = \begin{cases} 1, & \text{zooming} \\ 0, & \text{fixed camera} \\ -1, & \text{otherwise} \end{cases}. \quad (4)$$

Consequently, f_{cm} is positive during zooming, negative during translational camera and zero for fixed camera operations. The appearance of face and text also increases the attraction of such frames. The face detection possesses a high recognition value in key frames and pulls the viewers' attention. We use a trained Haar cascade classifier of OpenCV[1] to detect faces. The size and the number of faces give an explanation of the importance. A close-up is often more important than many small faces. A useful parameter is the area of faces F:

$$f_{face}(n) = \sum_{i=1}^{\mathrm{F}} f_{face_{width}}(i) \times f_{face_{height}}(i). \quad (5)$$

The parameter f_{text} is a binary value according to the text detection[2]:

$$f_{text}(n) = \begin{cases} 1, & \text{text detected} \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

Based on our analysis, we propose a method to integrate visual attractive elements into key frame extraction. In order to avoid inexpressive and visually similar key frames, we further integrate the normalised variance of luminance $f_{Y_{\sigma^2}}$ and the difference on Colour Layout [1] $f_{\mathrm{diff}_{CL}}$ (Eq. 7) to the last extracted key frame $f_{cl}(n = n_{lk})$.

$$f_{\mathrm{diff}_{CL}}(n) = \sqrt{\sum_{k=0}^{K-1} \left[f_{cl(k)}(n) - f_{cl(k)}(n_{lk})\right]} \quad (7)$$

A key frame is detected at the global maximum of the fusion curve f_{att} (Eq. (8)) in a (sub)shot according to Fig. 4:

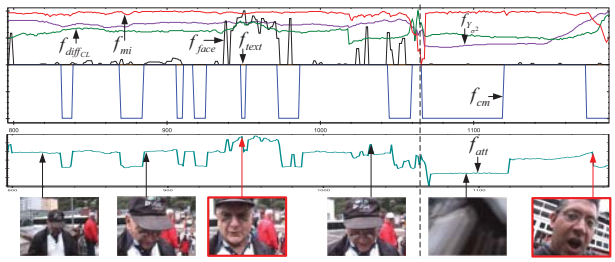

Fig. 4. Plots of single features (top) and fusion curve f_{att} (bottom); key frames (red boxed) are extracted at the maximum of f_{att} per shot

[1]Open Source Computer Vision Library by Intel
[2]Open Source OCR TessNet2

$$f_{att}(n) = w_{mi} \cdot f_{mi}(n) + w_{cm} \cdot f_{cm}(n) + w_{text} \cdot f_{text}(n)$$

$$+w_{face} \cdot f_{face}(n) + w_{Y_{\sigma 2}} \cdot f_{Y_{\sigma 2}}(n) + w_{\text{diff}_{cl}} \cdot f_{\text{diff}_{CL}}(n). \quad (8)$$

This fusion curve combines weighted attention features like motion intensity, special camera operations, text appearance and the size of detected faces. As all features are coequal, the weights w have the meaning of scaling. The weights can be changed, in order to emphasise a certain feature.

3. RESULTS & CONCLUSION

In this paper we have introduced a framework for segmenting user generated videos on video sharing websites (e. g. YouTube). There, we focus on videos on the unstructured channel "Travel". We compare our key frame extraction approach against the key frame extraction of IBM Multimedia Analysis and Retrieval System (IMARS)[3] (Fig. 5). Their key frame extraction is only based on visual differences. Many extracted key frames by IMARS are shaky, blurred and extracted during gradual transitions due to ignoring motion and visual attention features. Unlike IMARS, our approach extracts the key frame with the highest amount of visual attention and minimal motion intensity to get the steadiest frame. The key frames are evaluated by the mean opinion score (MOS), because it is hard to judge objectively the quality. The MOS is generated by averaging the ratings of 12 viewers for 18 travel videos (table 1). Our approach gets on the average a higher score. The key frames extracted by our approach get a score of 3.62 and the key frames by IMARS get a score of 2.99.

video	total frames	# shots	IMARS			our approach		
			# key frames	during transition	MOS	# key frames	during transition	MOS
New Zealand	3576	79	7	0	3	39	0	3.92
landing	930	1	1	0	2	1	0	1.92
kayaking	18343	81	46	1	3.25	75	0	3.67
zoo	3088	1	6	0	2.75	1	0	2.17
safari	6442	95	35	0	3.58	68	0	4
rainforest	18798	17	14	1	3.25	6	0	3.67
aurora	3172	20	3	0	2.08	15	0	3.75
autumn	8990	52	30	11	2.75	27	1	4.25
bridge Seoul	2430	110	6	0	3.33	26	0	3.33
Tibet	6323	39	17	0	3	20	0	2.92
Iran	7464	62	28	3	3.42	46	0	4.42
Chernobyl	12500	70	34	0	3.17	40	0	3.75
Mexico	3537	46	12	4	2.83	29	0	3.92
Colombia	8249	78	15	1	3.25	41	0	3.58
Peru	9022	89	29	5	3.17	32	0	3.83
Bratislava	1797	39	11	0	2.92	16	0	3.75
beach	631	1	3	0	2.8	5	0	4.2
Taj Mahal	9626	81	13	3	3.33	59	0	4.08
Results	*124918*	*961*	*310*	*29*	*2.99*	*546*	*0*	*3.62*

Table 1. The key frames, extracted by IMARS and by our approach, are evaluated by mean opinion score (MOS)

Future work will focus on the use of clustering methods to perform video similarity search on visual and textual features.

[3] http://www.alphaworks.ibm.com/tech/imars/

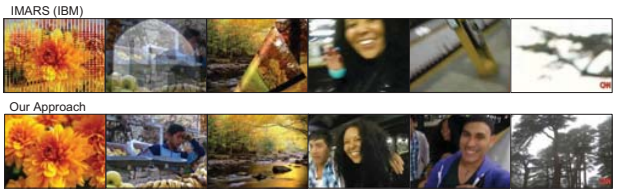

Fig. 5. Key frames extracted by IMARS (top); by our approach (bottom)

4. REFERENCES

[1] T. Sikora, P. Salembier, and B. S. Manjunath, *Introduction to MPEG-7: Multimedia Content Description Interface*, Number ISBN 047148678. John Wiley LTD, 2002.

[2] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation. a review," *Signal Processing Magazine*, vol. 23, no. 2, pp. 28–37, 2006.

[3] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Storage and Retrieval for Image and Video Databases*, 1999, vol. 3656, pp. 290–301.

[4] E. Saez, JI Benavides, and N. Guil, "Reliable real time scene change detection in MPEG compressed video," in *International Conference on Multimedia and Expo*, 2004, vol. 1.

[5] J. U. Won, Y. S. Chung, I. S. Kim, J. G. Choi, and K.H. Park, "Correlation based video-dissolve detection," in *Information Technology: Research and Education, 2003.*, pp. 104–107.

[6] X. Qian, G. Liu, and R. Su, "Effective Fades and Flashlight Detection Based on Accumulating Histogram Difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 10, pp. 1245, 2006.

[7] Z. C. Zhao and A. N. Cai, "Extraction of Semantic Keyframes Based on Visual Attention and Affective Models," in *Computational Intelligence and Security, 2007 International Conference on*, 2007, pp. 371–375.

[8] W. Chen and Y. J. Zhang, "Video segmentation and key frame extraction with parametric model," in *International Symposium on Communications, Control and Signal Processing.*, 2008, pp. 1020–1023.

[9] Z. Sun, K. Jia, and H. Chen, "Video Key Frame Extraction Based on Spatial-Temporal Color Distribution," in *Conference on Intelligent Information Hiding and Multimedia Signal Processing.*, 2008, pp. 196–199.

[10] Z. Cernekov, I. Pitas, and C. Nikou, "Information Theory-Based Shot Cut/Fade Detection and Video Summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, 2006.